Procesamiento del Lenguaje Natural, Revista nº 62, marzo de 2019, pp. 61-68 recibido 22-11-2018 revisado 15-01-2019 aceptado 10-02-2019

# Detección de plagio translingüe con grafos semánticos: experimentando con recursos en abierto

## *Detection of translingual plagiarism with semantic graphs: experimenting with open resources*

**Ana García-Serrano, Antonio Menta Garuz**
UNED, Universidad Nacional de Educación a Distancia
ETSI Informática, C/ Juan del Rosal 16, 28040 Madrid
agarcia@lsi.uned.es, mentared@gmail.com

**Resumen:** Hoy en día el idioma ha dejado de ser una barrera para plagiar documentos disponibles en Internet. Tras enfoques probabilísticos ya clásicos que no alcanzan buenos resultados con documentos multilingües con paráfrasis (Barrón-Cedeño, 2012), aparecen trabajos que, utilizando grafos de conocimiento, aumentan la capacidad semántica del análisis de las oraciones y mejoran los resultados de detección de plagio. Además, actualmente hay recursos lingüísticos, basados en el conocimiento, o de desarrollo de software que están disponibles para la experimentación, una vez decidido cuál de ellos elegir, cuáles están realmente disponibles en abierto, qué eficiencia aportan si se integran en la experimentación planteada, o qué tipo de características debe tener el ordenador o el servidor necesario para la investigación. Este trabajo plantea una investigación experimental para la detección de plagio translingüe siguiendo una línea de investigación y utilizando recursos disponibles en abierto. Los resultados alcanzan el estado del arte, y esperamos que el planteamiento seguido, el análisis justificado y las dificultades técnicas reportadas, acercará a los lectores la metodología necesaria en este tipo de experimentaciones y permitirá planificar sus trabajos futuros. El software desarrollado está disponible en abierto.
**Palabras clave:** Plagio translingüe, recursos lingüísticos, recursos en la red, experimentación, desarrollo de software

**Abstract:** Today the language has ceased to be a barrier to plagiarize documents available on the Internet. After classic probabilistic approaches that do not achieve good results with multilingual documents with paraphrasing (Barrón-Cedeño, 2012), there are works that, using knowledge graphs, increase the semantic ability in the analysis of sentences and improve the results of plagiarism detection. In addition, currently in linguistic engineering there are linguistic or knowledge-based resources, or software development resources that are available to experimentation once decided, which ones to choose, which ones are available, what efficiency they provide if they are integrated into the proposed experimentation, or what kind of features the computer or server should have to the investigation. This work proposes an experimental investigation into a concrete problem, the detection of translingual plagiarism following a line of research and using open resources. The results reach the state of the art, and we hope that the followed approach, the justified analysis and the technical difficulties reported, will bring readers closer to the methodology needed in this type of experimentation and will allow planning their future works. The software developed is available in open.
**Keywords:** Translingual plagiarism, linguistic resources, linked data, experimentation, software development

ISSN 1135-5948. DOI 10.26342/2019-62-7 

## 1 Introducción

Hay una diferencia importante entre inspirarse en obras de terceros y copiar el contenido intencionadamente, *"Plagiar es reusar las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y a su autor"* (Barrón-Cedeño et al., 2013).

Comas y Sureda (2008) constatan que el 61% de los universitarios españoles confiesa haber copiado de internet y el 3,3% incluso haber comprado documentos. Diez años después, en el mundo la proporción ha aumentado al 85% en estudiantes (Eaton et al., 2017) y el idioma ha dejado de ser una barrera. Son necesarias herramientas automáticas capaces de detectar los posibles casos de plagio, aunque la decisión final del mismo debería ser tomada por expertos en la materia.

Tras enfoques probabilísticos ya clásicos como la utilización de n-gramas a nivel de carácter o de corpus paralelos para cada idioma que no alcanzan buenos resultados con documentos multilingües con paráfrasis (Barrón-Cedeño, 2012), aparecen trabajos que, utilizando grafos de conocimiento, aumentan la capacidad semántica en el análisis de las oraciones y mejoran los resultados de detección de plagio. Además, actualmente en ingeniería lingüística hay un gran número de recursos lingüísticos, basados en el conocimiento *Linked Data* (LD) y recursos para desarrollo de software, que están disponibles para llevar a cabo las tareas de experimentación exigibles en este campo de investigación.

Pero hay preguntas que resolver a lo largo del proceso de desarrollo, como cuál elegir, cuáles están realmente disponibles, qué eficiencia aportan si se integran en la experimentación planteada, o qué tipo de características debe tener el ordenador o servidor, y todas ellas, sin conocer si los resultados serán los esperados.

Los objetivos de este artículo son tanto avanzar en la investigación de la detección translingüe, como mostrar las decisiones que exige una experimentación científica usando recursos en abierto.

En lo que sigue, se identifican los tipos de plagio, se presenta una breve revisión del estado del arte, así como la serie de recursos PAN y algunas herramientas automáticas actuales para detección de plagio, que muestran que el problema sigue abierto.

A continuación, siguiendo la línea de investigación realizada por Franco-Salvador et al. (2016a) y Franco-Salvador (2017), se describe la propuesta de detección de plagio translingüe utilizando recursos disponibles como son *Freeling* para el análisis lingüístico de textos multilingües, *BabelNet*, un diccionario semántico multilingüe que promete abstraer del idioma de los conceptos que aparecen en los textos, o recursos de software como *GraphStream*, para la gestión de grafos de conocimiento (Menta, 2018).

Finalmente se incluyen los detalles de las pruebas y resultados obtenidos con el prototipo desarrollado y las conclusiones.

## 2 Trabajos relacionados

La detección de plagio admite clasificaciones desde diferentes puntos de vista. En (Martin, 2004) se clasifican según el objetivo del plagio: (1) plagio de ideas; (2) plagio palabra por palabra (sin el texto con comillas); (3) plagio de fuentes y (4) plagio de autoría.

El estado del arte también se puede clasificar según las aproximaciones utilizadas, orientadas al estudio de características internas del documento (longitud de palabras, frecuencia de uso, número de adjetivos, etc.) o al estudio de características externas (cálculo de la similitud con fragmentos de terceros, distribución similar de palabras, etc.) (Meyer et al., 2007). El primer tipo se conoce como detección intrínseca y el segundo como detección externa, al recurrir a un conjunto de documentos externos.

O bien se pueden clasificar según el modo de copia, exacta o modificada (paráfrasis) (Barrón-Cedeño et al., 2013). En el segundo caso se modifica el texto sustituyendo ciertas partes con significado similar o eliminando algún elemento, y la detección se centra en la semejanza entre fragmentos de texto.

En Franco-Salvador et al. (2012 y 2016b), se clasifican los enfoques para detección de plagio translingüe en modelos basados en: (1) diccionarios, reglas y tesauros lingüísticos que traducen de forma aislada conceptos y palabras; (2) la sintaxis de cada documento; (3) corpus comparables, con documentos en diferentes idiomas que describen de forma aproximada el mismo contenido y (4) corpus paralelos que contienen documentos con el contenido exacto en diferentes idiomas.

Este trabajo de detección de plagio translingüe (EN-ES) es clasificable como detección de plagio de ideas, externa, de copia modificada o paráfrasis y detección semántica de plagio utilizando un diccionario enciclopédico del LD que permite comparar los documentos sospechosos con los del corpus.

## 2.1 Herramientas para la detección automática de plagio

La diferencia en la calidad, cobertura y precio de las herramientas automáticas es grande, aunque el modelo de negocio sea muy similar, y depende de la eficiencia y la calidad del corpus. En la mayoría de ellas se desconocen las técnicas que se utilizan para descubrir el plagio.

En Nahas (2017) se describen varias herramientas en la red, tanto gratuitas como de pago, que se comparan con los siguientes aspectos: (As1) Utilizable para la prevención de casos de plagio académicos; (As2) Utilizable sin necesidad de ser descargada por el usuario final (integrada en los servidores de correo, página web y otras); (As3) Corpus de internet o/y propio (como otros trabajos académicos) y (As4) Aporta otras funcionalidades.

Los detalles y comentarios de las cinco herramientas seleccionadas son los siguientes:

**Urkund**[1]. Es As1, As2, As3. Actualmente disponen de más de 23 millones de documentos.

**Turnitin**[2]. Es As1, As3, As4. La aplicación busca coincidencias entre 61.000 millones de páginas indexadas, más de 600 millones de trabajos de estudiantes y 150 millones de artículos, libros y periódicos. Destaca en la buena detección de copia exacta (analiza grupos de ocho a diez palabras). Prácticamente es la única herramienta capaz de detectar plagio translingüe, traduciendo los contenidos al inglés y realizando detección monolingüe en inglés.

**Unplag**[3]. Es As1, As3, As4. Centrado en trabajos académicos, realiza una búsqueda de fragmentos del documento con los índices web de Bing y Yahoo. Permite al usuario analizar hasta cinco documentos de forma simultánea.

**PlagiarismCheck**[4] Es As1, As3, As4. Muy efectiva tanto para académicos como para propietarios de sitios web, redactores, bloggers y otros, y revisa los textos web para ver si se roban en línea. Indica que detecta todos los tipos de plagio.

**Plagiarisma.** Es As1, As2, As3, As4. Página web que permite subir archivos o pegar directamente el texto (max. de 1000 caracteres en cada prueba) (Krizkova et al., 2016).

Por lo tanto, la detección de plagio translingüe con paráfrasis es un tema abierto también entre la mayoría de las herramientas comerciales.

## 2.2 Recursos PAN

La iniciativa PAN (*Uncovering Plagiarism Authorship and Social Software Misuse*), es un referente internacional en el área y se creó en 2007 como evento científico para promover la investigación forense de textos digitales. En 2010 se desarrolló un *framework*, que facilita la reproducibilidad, al incluir un corpus y una métrica de evaluación (Potthast et al., 2010 y 2011b).

En la primera edición se desarrolló un corpus basado en un conjunto de libros del proyecto Gutenberg, grandes trabajos de la literatura clásica, la mayoría en inglés. Como estos libros no contienen casos de plagio conocido, se crearon manual y artificialmente los casos de plagio. Del total de libros del proyecto Gutenberg, el corpus PAN-PC-10 contiene 22.000 libros en inglés, 520 en alemán y 210 en español.

En la segunda edición de la competición se mejoró el corpus con la inclusión de fragmentos sospechosos en documentos con temas similares, en vez de sólo incluirlos de forma aleatoria en cualquier tipo de documento. Para hacerlo, el contenido del corpus PAN-PC-10 se dividió en clústeres según temas y la mitad de los fragmentos artificiales creados son insertados en el mismo clúster. También se añadieron nuevos tipos de plagio con el fin de aumentar el realismo del corpus. La novedad en 2011 fue la inclusión de casos basados en paráfrasis, porque en la mayoría aparece algún tipo de modificación en la estructura o en los elementos que forman las frases, y sobre todo cuando el plagio es translingüe. En Potthast et al. (2011a) se resume el contenido del corpus PAN-PC-11, con 26.939 documentos y un total de 61.064 casos de plagio, de los que cerca del 70% presenta paráfrasis.

Muchas de las investigaciones desde entonces han utilizado la versión PAN-PC-10 del corpus para la parte de desarrollo de su

[1] https://www.urkund.com/es/
[2] http://www.turnitin.com/
[3] https://es.unplag.com/
[4] https://plagiarismcheck.org

aplicación y PAN-PC-11 para comprobar los resultados, como es el caso de este trabajo.

En la actualidad para temas como la ciberseguridad, análisis de redes, forense y otras, además de la detección de plagio también se investiga en la detección de autoría de un texto (Kestemont et al., 2018), aunque queda fuera del objeto de este trabajo.

## 2.3 Técnicas de detección automática de plagio translingüe

El modelo *Cross-language character n-gram* (CL-CNG) utiliza n-gramas a nivel de caracteres para dividir los documentos en fragmentos comparables. Una vez normalizados los términos, es habitual representar el documento como un vector con todas las posibilidades de aparición conjunta de los elementos del alfabeto ($M_c$Namee et al., 2004).

El modelo *Cross-language alignment-based similarity* (CL-ASA) utiliza un corpus paralelo de textos (Barrón-Cedeño, 2012). Cada idioma tiene un factor de longitud diferente, determinado por la media y la desviación estándar de la longitud de los caracteres de una traducción de un idioma a otro (Tabla 1).

| Para-meter | en-de | en-es | en-fr | en-nl | en-pl |
|---|---|---|---|---|---|
| μ | 1.089 | 1.138 | 1.093 | 1.143 | 1.216 |
| σ | 0.268 | 0.631 | 0.157 | 1.885 | 6.399 |

Tabla 1 - Factor de longitud

El elemento fundamental de este modelo es el diccionario estadístico para la traducción, que incluye la probabilidad de traducción de una palabra de un idioma a otro para calcular la similitud entre dos documentos, mostrando un elevado rendimiento a bajo coste computacional (Franco-Salvador et al., 2012).

En Franco-Salvador et al. (2013) se describe la aproximación *Cross-language knowledge graphs analysis* (CL-KGA), utilizando grafos de conocimiento, donde los nodos son conceptos del documento, las aristas unen dos conceptos relacionados, y el peso de la arista muestra la importancia de la relación. Una vez obtenidos los grafos de los documentos, sospechoso y originales, se aplica una medida de la similitud entre ellos.

La técnica *Plagiarism detection using linguistic knowledge* (PLDK) combina la información sintáctica con la semántica entre los conceptos del documento (Abdi et al., 2015). Calcula la similitud a partir de un vector de similitud en el orden de las palabras y otro de similitud semántica mediante consultas a *WordNet*, buscando el ancestro común y su distancia, que servirá de índice de semejanza.

En Franco-Salvador et al. (2016a) se propone una aproximación hibrida: utilizar grafos de conocimiento para calcular la similitud semántica por medio de recursos de redes semánticas como *WordNet* o *BabelNet*, y un modelo basado en el espacio vectorial para capturar los aspectos sintácticos que las redes semánticas no son capaces de detectar.

En los últimos años se ha producido un giro hacia las técnicas basadas en *Deep Learning*, frecuentemente utilizando una representación de palabras basada en *word embeddings*, para aportar conocimiento semántico (Suleiman et al. 2017), (Gupta, 2017).

# *3 Propuesta*

El objetivo de este trabajo es encontrar una solución al problema de detección de plagio translingüe, partiendo de la técnica CL-KGA, utilizando herramientas disponibles de *Linked Data*, porque (1) la utilización de grafos de conocimiento permite abstraer tanto el problema del idioma a la hora de calcular la semejanza de dos documentos, como del orden de aparición de las palabras en las oraciones (clave para resolver paráfrasis) y (2) pueden utilizarse recursos semánticos para obtener otra información semántica del contenido.

Como se detalla en los apartados siguientes, esta propuesta exige: (1) La selección del corpus, PAN-PC-10, para entrenar el modelo en busca de los mejores parámetros posibles, y PAN-PC-11 para la fase de prueba, (2) el procesado lingüístico de los documentos y (3) la selección de las métricas de similitud entre los grafos de conocimiento y entre los documentos.

## 3.1 Procesamiento lingüístico

Las tareas de procesamiento de lenguaje necesarias se realizan con *Freeling*[5] y son: detección de idioma, tokenización del documento, división del documento en oraciones, análisis y etiquetación morfológica, análisis sintáctico de la oración y lematización.

Se selecciona *Freeling* por las funcionalidades que incluye en varios idiomas,

[5] http://nlp.lsi.upc.edu/freeling/node/1

sus diccionarios morfológicos de calidad, y porque puede integrarse con otras aplicaciones.

Como recurso semántico se ha escogido *BabelNet*[6] porque es un diccionario enciclopédico multilingüe con más de 14 millones de entradas que conecta entidades y conceptos en 271 idiomas (v. 3.7). Cada entrada o *synset*, representa un sentido de un concepto y contiene sus sinónimos.

Para seleccionar los fragmentos de los documentos, siguiendo a Franco-Salvador et al. (2016a), por defecto son de 5 sentencias consecutivas y con un salto posterior de 2 sentencias. Se preprocesan las palabras del fragmento para el formato (palabra_lema, etiqueta) de consulta en *BabelNet*, y la información extraída se organiza en grafos.

Para conseguir todos los sentidos de cada uno de los fragmentos de texto, que son los nodos iniciales del grafo, se piden los *synsets* vecinos y se estudia su coincidencia con algún synset ya almacenado como nodo, con la excepción de aquellos que pertenecen a la misma palabra. En caso de encontrar otro *synset* existente en el grafo se añade la relación entre los dos nodos. Este proceso es recursivo hasta un número máximo de saltos (parámetro).

## 3.2 Similitud entre grafos de conocimiento

El cálculo de la similitud entre dos grafos, G y G', en este trabajo se basa en los nodos comunes y en el número de aristas que los une. Se define por $S_c(G,G')$, basada en el coeficiente de Dice entre los pesos de los nodos.

$$S_c(G, G') = \frac{2 \cdot \sum_{c \in V(G) \cap V(G')} w(c)}{\sum_{c \in V(G)} w(c) + \sum_{c \in V(G')} w(c)},$$

Donde V(G) y V(G') son los dos grafos por comparar y w(c) es el número de aristas incidentes en el vértice. El valor de semejanza entre dos fragmentos se encuentra en [0, 1]. En caso de que los dos grafos contengan sentidos y conceptos similares, el grafo de intersección entre los dos contendrá un alto número de nodos respecto al grafo unión. Por lo tanto, la medida se alejará de 0 y estará más cercana a 1. Por el contrario, si el grafo de intersección está vacío, será 0. Cada vez que la comparación de dos fragmentos supere un valor umbral (que se define de forma empírica), se marca como un resultado positivo (parcial) de ser plagio. En caso de obtenerse un resultado positivo en tres fragmentos consecutivos del documento, el algoritmo confirma el plagio.

## 3.3 Métrica de evaluación

La métrica para evaluar el resultado de la detección de casos de plagio es el *F1-score* en este trabajo, aunque en las competiciones PAN se utiliza *plagdet*, que también se basa en la *precisión*, *exhaustividad* y, además, en la granularidad de casos positivos.

$$plagdet(S, R) = \frac{F_\alpha}{\log_2(1 + gran(S, R))}.$$

Donde S es el conjunto de casos reales de plagio, R es el conjunto de detecciones, gran(S, R) es un valor de granularidad (el número de casos positivos de un mismo plagio) y Fα es el valor *F1-score*. *F1-score* se diferencia de *pladget* en la granularidad, para evitar redundancia. Como en este trabajo solo se tiene en cuenta un caso positivo, *plagdet* se convierte en el valor de *F1-score*.

## 3.4 Diferencias

Hay cuatro diferencias fundamentales entre la aproximación descrita en Franco-Salvador et al. (2016a y 2016b) y la presentada: (1) Respecto a la métrica de similitud entre grafos, en este trabajo se utiliza el grado incidente en los vértices y en el suyo una interpolación entre aristas y vértices; (2) Para la generación de los grafos de conocimiento, se exige un máx. de tres saltos y ellos reinician el número de saltos al encuentrar un vértice; (3) Para determinar plagio, son suficientes 3 grafos consecutivos positivos y ellos, para cada fragmento eligen los 5 más parecidos y aplican una función de convergencia uniendo fragmentos hasta superar un límite; (4) La hipótesis de trabajo de utilizar granularidad uno, hace que la métrica *plagdet* se convierta en *F1-score*.

# *4 Prototipo experimental*

Descargados los dos corpus PAN[7], el prototipo se encarga de crear los fragmentos de texto, el procesado lingüístico con *Freeling*, asociar a cada fragmento un grafo de conocimiento utilizando *BabelNet*, la serialización a disco y la recuperación de cada documento con sus fragmentos y grafos de conocimiento asociados,

[6] https://babelnet.org/

[7] http://pan.webis.de/

calcular la similitud entre grafos y de la decisión de plagio cuando en la comparación de fragmentos haya tres resultados positivos.

### 4.1 Otros detalles técnicos

Se ha elegido la librería de grafos *GraphStream*[8], por su facilidad de uso, la disponibilidad de los algoritmos más habituales, la capacidad dinámica para añadir tanto nodos como relaciones, su visualización de los grafos y la capacidad de serialización de la librería.

Debido al alto consumo de memoria al generar los grafos, el preproceso con *Freeling* se ha desplegado en un Ubuntu 14.04 LTS, con un wrapper Java alrededor de la aplicación C++ y con JAX-RS se ha obtenido un endpoint de consulta (Apache 7.0)[9]. Para una máquina diferente, modificar la *url* y *localhost* por su *ip*.

El prototipo se encuentra disponible en https://github.com/Hisarlik/CrossLanguagePlagiarism/.

### 4.2 Dificultades integrando recursos semánticos

Las principales dificultades relacionadas con el uso de recursos en abierto han sido:

**Instalación y utilización de *Freeling*,** que es laboriosa por la gran cantidad de librerías de C a compilar y las variables de entorno y sistema necesarias (cada distribución Linux tiene diferentes configuraciones de carpetas).

**Configuración librerías Java,** para acceder a las funcionalidades de *Freeling*, en cuyo repositorio de github hay código para un API Java. Para este trabajo se desarrolló una aplicación API Rest compleja, para (por ejm.) referenciar las funcionalidades por separado.

Modificación de los **Timeouts en la API Rest** desarrollada para el procesado de los documentos especialmente largos.

**Instalación de *BabelNet*.** La versión 4.0 dispone de una API para unas pocas peticiones. Gracias a la política para investigación del startup babelscape.com, se descargó en local la base de datos (16GB archivos Lucene).

**Uso de memoria de los grafos.** Además del tiempo de computación, como casi todos los grafos pasan de mil nodos y el prototipo está en un Mac de 16GB de RAM, de los cuales 4GB dedican a *Freeling*, si se guardaban los grafos en memoria, se quedaba sin ella, y se decidió serializar con *GraphStream* (interfaz en Java, de *Serializable* para las clases).

**Lógica recursiva de búsqueda de los nodos del grafo.** La mayor dificultad debida a la capacidad del equipo personal usado ha sido implementar la búsqueda de relaciones entre los nodos iniciales del grafo y su expansión con *synsets* intermedios entre ellos. En un equipo como el descrito, el tiempo de creación de un grafo suele ser de 5m y procesar un documento 100m, así que en un tiempo razonable solo era asumible probar en un subconjunto del corpus.

## *5 Pruebas y evaluación*

Una vez desarrollado el prototipo, hay que configurar los parámetros del algoritmo para documentos en español e inglés. Estos son: (P1) Profundidad de peticiones a *BabelNet* o número de saltos permitidos para encontrar una relación entre dos conceptos; (P2) Umbral de similitud entre fragmentos y (P3) Número de fragmentos similares consecutivos para identificar plagio.

### 5.1 Pruebas con PAN-PC-10 y 11

La primera prueba sobre un subconjunto monolingüe (inglés) con 96 documentos sospechosos, 35 originales y 35 casos reales de plagio obtuvo el mejor resultado con la configuración: (P1) 2, (P2) 0.35 y (P3) 3. Se detectan 20 posibles casos de plagio, de los cuales 9 falsos positivos, con 24 casos no detectados. El algoritmo alcanzó Precisión 0.55; Exhaustividad (o *recall*) 0.31 y *F1-score* 0.4.

Los mejores resultados de todas las pruebas realizadas (Tabla 2), muestran que la ampliación del número de grafos consecutivos para decidir plagio no mejora la detección, y con valores de similitud entre grafos mayor que 0.3, aumentan los falsos positivos.

| Configuración | Casos Reales Detect | Detec-ciones Total | Casos Reales Total | Preci-sión | Recall | F1-score |
|---|---|---|---|---|---|---|
| sim > 0.3 y 3 frag. consecuti. | 15 | 78 | 35 | 0.19 | 0.43 | 0.27 |
| sim > 0.4 y 3 frag. consecu. | 8 | 9 | 35 | 0.89 | 0.23 | 0.36 |
| sim > 0.3 y 4 frag. consecu. | 9 | 30 | 35 | 0.3 | 0.26 | 0.28 |
| sim> 0.35 y 4 frag. cons. | 9 | 12 | 35 | 0.75 | 0.26 | 0.38 |
| sim> 0.35 y 3 frag. cons. | 11 | 20 | 35 | 0.55 | 0.31 | 0.40 |

Tabla 2 – Resultado: solo inglés

[8] http://graphstream-project.org/

[9] https://github.com/Hisarlik/CrossLanguagePlagiarism_API_Freeling

Las siguientes pruebas fueron para detección multilingüe, entre textos en español e inglés. El subconjunto del corpus tiene 35 documentos sospechosos en inglés, 40 originales en español y 47 casos reales de plagio.

El mejor resultado se ha obtenido con la configuración: (P1) 2, (P2) 0.32 y (P3) 3. Se han detectado 32 posibles casos de plagio de los cuales 28 lo eran y 4 no y los otros 19 casos posibles no han sido detectados. El algoritmo alcanza una Precisión 0.875; Exhaustividad 0.595 y *F1-score* 0.708. La principal conclusión es que se identifican pocos falsos positivos.

Aparte de este resultado, se han realizado otras pruebas cuyos mejores resultados se muestran en la Tabla 3, concluyéndose que si el valor de similitud es 0.3 o mayor, el número de falsos positivos aumenta, pero los resultados generales mejoran.

| Configuración | Precisión | Recall | F1-score |
|---|---|---|---|
| **similitud > 0.3 y 3 grafos consecutivos** | 0.53 | 0.49 | 0.58 |
| **similitud > 0.35 y 3 grafos consecutivos** | 0.96 | 0.66 | 0.65 |
| **similitud > 0.25 y 4 grafos consecutivos** | 0.12 | 0.68 | 0.20 |
| **similitud 0.32 y 3 grafos consecutivos** | 0.87 | 0.59 | 0.70 |

Tabla 3 – Resultado español - inglés

Una vez encontrados los parámetros óptimos, estos se han utilizado para procesar el corpus PAN-PC-11 (34 documentos sospechosos en inglés, 33 originales en español y 40 casos reales). De los 40 casos reales de plagio, en la mejor prueba se detectan 16, con 7 falsos positivos, alcanzando una Precisión 0.70, *Recall* 0.40 y F1-score 0.51, resultados inferiores a las pruebas con PAN-PC-10, mostrando la dificultad del plagio translingüe con paráfrasis.

## 5.2 Comparación con otros trabajos

Antes de comparar los resultados con los de las dos competiciones PAN, es necesario recordar que, en este trabajo (1) no se ha utilizado la totalidad del corpus, solo aquellos casos en que los documentos originales están en español y los sospechosos en inglés y (2) se utiliza el *F1-score* porque el algoritmo propuesto es de granularidad 1, recordando que algunos de los trabajos presentados también utilizan esta granularidad.

Los mejores resultados de los 11 trabajos en PAN 2010, publicados en[10], son: *pladget* 0,79; 0,70; 0,69. El mejor en este trabajo es *F1-score* 0,70. Los mejores resultados oficiales de los 9 de PAN 2011, publicados en[11], son: pladget 0,55; 0,41; 0,34. Comparando con el enfoque CL-KGA (DCW) con pladget 0.65 y CL-KGA (WSD concepts y/o weighting) con pladget 0.64, los resultados obtenidos son inferiores a estos últimos. El mejor en este trabajo es F1-score 0,51, similar a los mejores obtenidos en la competición.

## *6 Conclusiones*

De acuerdo con los resultados obtenidos, se puede afirmar que la selección de un modelo basado en grafos de conocimiento y utilizando recursos semánticos (en abierto) para la detección de posibles casos de plagio translingüe (español e inglés), presenta resultados comparables con los obtenidos en las competiciones PAN 2010 y 2011.

Otro objetivo del trabajo era justificar la selección de recursos en abierto y cómo resolver las dificultades. Los resultados muestran la adecuación de los recursos para la gestión de los grafos, el procesado lingüístico y la aproximación seguida para definir los parámetros óptimos del algoritmo.

Sin embargo, es una solución costosa e intensiva en tiempo, memoria y CPU en un ordenador personal como el utilizado, por lo que en las pruebas solo se trabajó con documentos en inglés y español. En estas condiciones, se podría mejorar el rendimiento, ya sea paralelizando la creación de los grafos o guardando temporalmente los nodos que aparecen en la mayoría de los grafos.

Sería interesante estudiar modificaciones en el cálculo de la similitud entre los grafos, otras formas de ponderar la importancia de los conceptos para aumentar la calidad de los resultados o probar con otras aproximaciones para el problema planteado.

## *Agradecimientos*

Este trabajo ha sido parcialmente financiado por los proyectos Musacces (S2015/HUM3494) y VEMODALEN (TIN2015-71785-R).

[10] https://pan.webis.de/clef10/pan10-web/plagiarism-detection.html

[11] https://pan.webis.de/clef11/pan11-web/plagiarism-detection.html

## *Bibliografía*


Abdi, A., N. Idris, R. Aliguliyev y R. M. Aliguliyev. 2015. PDLK: Plagiarism detection using linguistic knowledge. *Expert Systems with Applications*, 42(22): 8936-8946.

Barrón-Cedeño, A. 2012. On the Mono and Cross-Language Detection of Text Re-Use and Plagiarim. *Ph.D. thesis*, DSIC, UPV.

Barrón-Cedeño, A., M. Vila y P. Rosso. 2013. Plagiarism meets Paraphrasing: Insights for the Next Generation in Autom. Plagiarism Detection. In: *Computational Linguistics*, 39(4) 917-947.

Comas R. y J. Sureda. 2008. Academic cyberplagiarism: tracing the causes to reach solutions. *The Humanities in the Digital Era*, 10:1-7.

Eaton S., M. Guglielmin y B. Otoo. 2017. PLAGIARISM: Moving from punitive to proactive approaches. In *Selected Proc. of the IDEAS Conference: Leading Educational Change*, páginas 28-36.

Franco-Salvador, M., P. Gupta y P. Rosso. 2012. Detección de plagio translingüe utilizando el diccionario estadístico de BabelNet. *Computación y Sistemas*, 16(4): 383-390.

Franco Salvador, M., P. Gupta y P. Rosso. 2013. Cross-language plagiarism detection using multilingual semantic network. *Proc. ECIR* Springer, páginas 710-713.

Franco-Salvador M., P. Gupta, P. Rosso y E. Banchs. 2016a. Cross-language plagiarism detection over continuous space and knowledge graph-based representations of language. *Knowledge-based systems* 111, páginas 87-99.

Franco-Salvador M., P. Rosso y M. Montes 2016b. A Systematic Study of Knowledge Graph Analysis for Cross-language Plagiarism Detection. *Information Processing & Management*, 52(4): 550–570.

Franco-Salvador M. 2017. A Cross-domain and Cross-language Knowledge-based Representation of Text and its Meaning. *Ph.D. thesis*, DSIC, UPV.

Gupta P. 2017. Cross-View Embeddings For Information Retrieval. *Ph.D. thesis*, DSIC, UPV.

Kestemont, M., M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein y M. Potthast. 2018. Overview of the Author Identification Task at PAN-2018 Cross-domain Authorship Attribution and Style Change Detection. *Proc. CLEF*, CEUR 2125.

Krizkova, S., H. Tomaskova y M. Gavalec. 2016. Preference comparison for plagiarism detection systems. Fuzzy Systems (FUZZ-IEEE), páginas 1760-1767.

Martin, B. 2004. Plagiarism: policy against cheating or policy for learning? Australia. https://ro.uow.edu.au/artspapers/78/.

McNamee, P. y J. Mayfield. 2004. Character n-Gram Tokenization for European Language Text Retrieval. *Information retrieval*, 7(1-2): 73-97.

Menta, A. 2018. Detección de plagio multilingüe mediante recursos semánticos. *Tesis de Máster*. ETSI Informática, UNED.

Meyer, S., B. Stein y M. Kulig. 2007. Plagiarism Detection without Reference Collections. In *Advances in data analysis*, Springer, páginas 359-366.

Nahas, M. 2017. Survey and Comparison between Plagiarism Detection Tools. *American J. of Data Mining and Knowledge Discovery*, 2(2): 50–53.

Potthast M., A. Barrón-Cedeño, B. Stein y P. Rosso. 2010. An Evaluation Framework for Plagiarism Detection. In *proc. COLING-2010*, páginas 997 -1005

Potthast M., A. Eiselt, A. Barrón-Cedeño, B. Stein y P. Rosso. 2011a. Overview of the 3rd International Competition on Plagiarism Detection. In: Petras V., Forner P., Clough P. (Eds.), *Notebook Papers of CLEF 2011 LABs and Workshops*. In CEUR workshop proceedings, Vol. 1177.

Potthast M., A. Barrón-Cedeño, B. Stein y P. Rosso. 2011b. Cross-Language Plagiarism Detection. In: *Languages Resources and Evaluation*. Special Issue on Plagiarism and Authorship Analysis, 45(1): 45-62.

Suleiman, D., A. Awajan y N. Al-Madi. 2017. Deep Learning Based Technique for Plagiarism Detection in Arabic Texts. In *New Trends in Computing Sciences* (ICTCS) IEEE, páginas 216-222.